

# A new implementation of relative distribution methods in Stata

Ben Jann

University of Bern

2020 Swiss Stata Conference  
University of Bern, November 19, 2020

# Outline

- 1 Introduction
- 2 Theory and estimation
- 3 The `reldist` command
- 4 Spinoff: a general command for the analysis of distributions

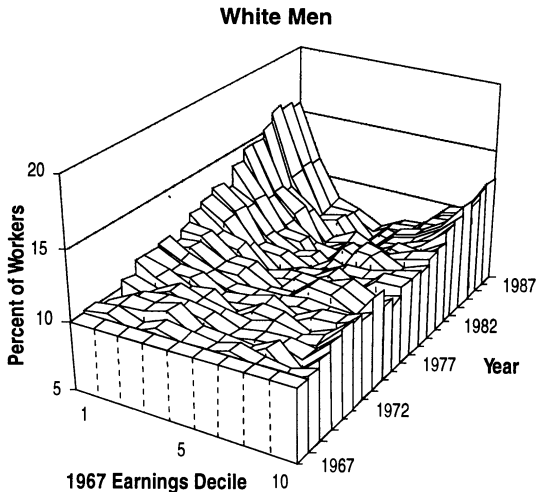
# What is the “relative distribution”?

- The relative distribution is the distribution of the relative ranks that the outcomes from one distribution take on in another distribution.
- How do wages of females rank in the wage distribution of males?  
How are these ranks distributed?
- The method can be used to analyze differences in distributions between groups or changes in a distribution over time.
- Of interest are aspects such as the distribution function or the density function of the relative ranks, or summary statistic such as polarization or distributional divergence.
- Of interest are also counterfactual decompositions that adjust the relative distribution for differences in covariate compositions.

# Example: Polarization of earnings over time

(Morris et al. 1994)

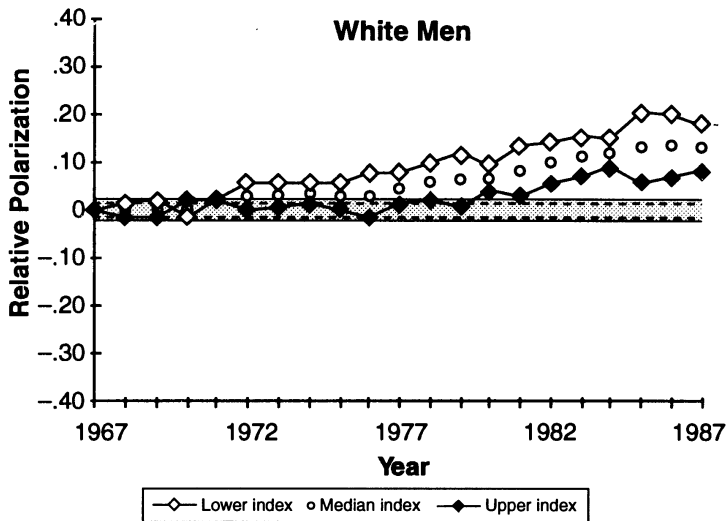
- Change in earnings of full-time, full-year workers: relative distribution of a given year compared to 1967



# Example: Polarization of earnings over time

(Morris et al. 1994)

- Relative earnings polarization with respect to 1967



- 1 Introduction
- 2 Theory and estimation
- 3 The `reldist` command
- 4 Spinoff: a general command for the analysis of distributions

# Some definitions

- $F_Y$ : reference distribution (wages of males)
- $F_X$ : comparison distribution (wages of females)
- Relative distribution

$$G(r) = F_X(F_Y^{-1}(r)), \quad r \in [0, 1]$$

- Relative density

$$g(r) = \frac{dG(r)}{dr} = \frac{f_X(F_Y^{-1}(r))}{f_Y(F_Y^{-1}(r))}, \quad r \in [0, 1]$$

- Relative ranks

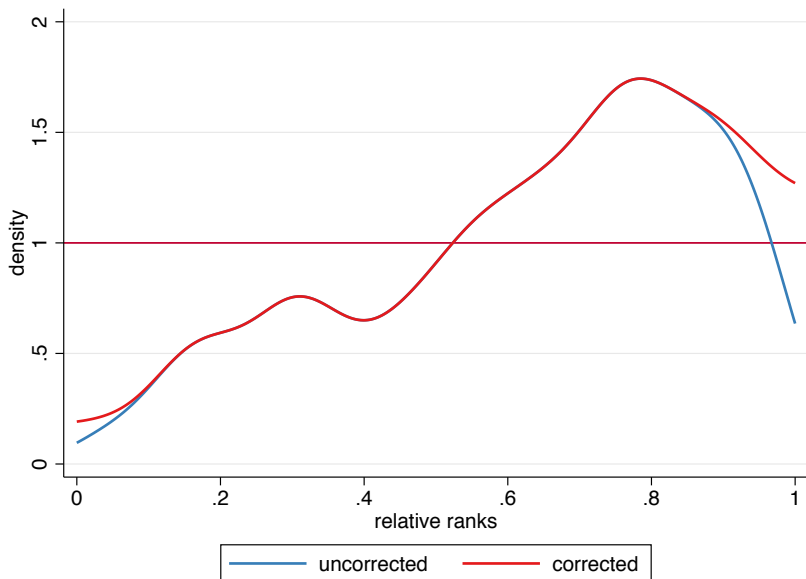
$$r_i = F_Y(X_i), \quad i \in \mathcal{X}$$

# Estimation

- Estimation of the relative CDF and summary measures of the relative ranks is pretty much straightforward.
- Estimation of the PDF is more involved:
  - ▶ Standard density estimators are (severely) biased at the boundaries because relative ranks can only take on values between 0 and 1.
  - ▶ Data-driven bandwidth selection requires adjustment to take account of the two-sample nature of relative data.
  - ▶ Function `mm_density()` from `moremata` can handle both issues.
- Estimation of standard errors is *not* straightforward due to the two-sample nature of the estimation problem.
  - ▶ I use influence functions based on an analogy to GMM (also see Jann 2020a).
  - ▶ The influence functions also cover uncertainty induced by covariate balancing.
  - ▶ Advantage of influence functions: Full support for complex survey estimation.



# Boundary effects



- 1 Introduction
- 2 Theory and estimation
- 3 The `reldist` command
- 4 Spinoff: a general command for the analysis of distributions

# The `reldist` command

- `reldist` provides a full-blown implementation of relative distribution methods.
  - ▶ Relative CDF and PDF for continuous and discrete data.
  - ▶ Relative polarization and divergence measures.
  - ▶ Summary statistics of relative ranks such as mean and quantiles.
  - ▶ Shape and location decomposition.
  - ▶ Covariate balancing by inverse probability weighting (IPW) or entropy balancing.
  - ▶ Utility to create graphs.
  - ▶ VCE for everything, including support for `svy` (although not as prefix command; must specify option `vce(svy)`)
  - ▶ Prediction of influence functions after estimation.
- For formulas and detailed information on the command see Jann (2020b).

## Estimation

Two-sample relative distribution (syntax 1)

```
reldist subcmd varname [if] [in] [weight], by(groupvar) [ options ]
```

Paired relative distribution (syntax 2)

```
reldist subcmd varname refvar [if] [in] [weight] [, options ]
```

where *subcmd* is

<b>pdf</b>	relative density
<b><u>histogram</u></b>	relative histogram
<b>cdf</b>	relative cumulative distribution
<b><u>divergence</u></b>	divergence measures
<b>mrp</b>	median relative polarization
<b><u>summarize</u></b>	summary statistics of relative ranks

Replay results

```
reldist [, noheader notable display_options ]
```

Draw graph after estimation

```
reldist graph [, graph_options ]
```

Obtain influence functions after estimation

```
predict {stub* | newvar1 newvar2 ...} [if] [in] [, scores density_options ]
```

# Example: Gender wage gap in Switzerland

```
. use sess16, clear
(Sample from Swiss Earnings Structure Survey 2016)
```

```
. describe
```

Contains data from sess16.dta

obs:	100,000	Sample from Swiss Earnings Structure Survey 2016
vars:	5	18 Nov 2020 19:02

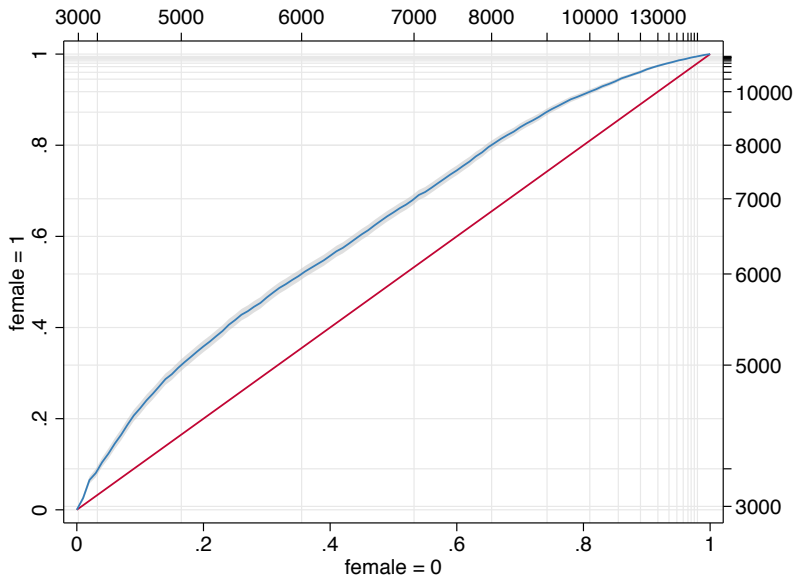
variable name	storage type	display format	value label	variable label
earnings	long	%10.0g		monthly earnings in CHF (full-time equivalent)
female	byte	%8.0g		1 = female, 0 = male
educyrs	byte	%10.0g		years of education
tenure	byte	%8.0g		tenure (in years)
wgt	double	%10.0g		sampling weight

Sorted by:

```
. summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
earnings	100,000	7858.498	4249.54	2312	103998
female	100,000	.44628	.4971083	0	1
educyrs	100,000	12.67786	2.728897	7	17
tenure	100,000	8.57528	8.905727	0	61
wgt	100,000	33.13712	59.26461	8.435029	2991.433

# Relative CDF



# Relative CDF



```
. reldist cdf earnings [pw=wt], by(female) notable
```

```
Cumulative relative distribution
```

```
Number of obs      =      100,000
```

```
F1: female = 1
```

```
Comparison obs     =      44,628
```

```
F0: female = 0
```

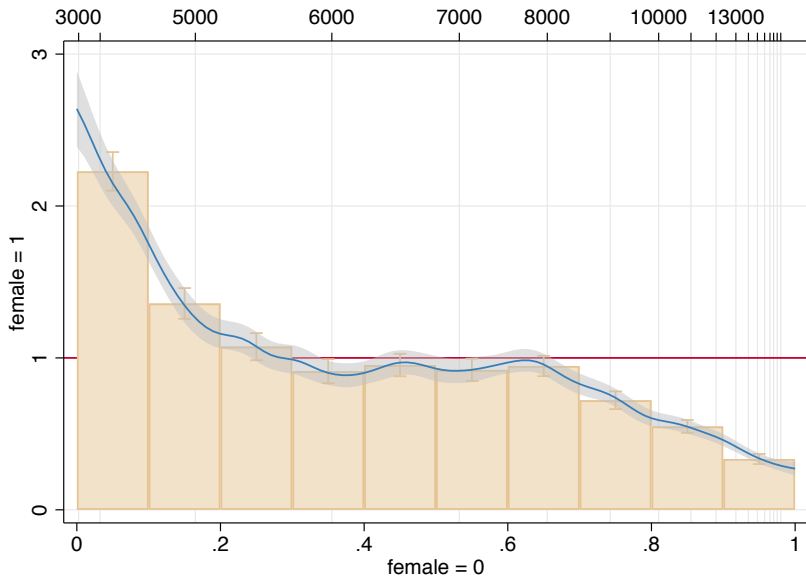
```
Reference obs      =      55,372
```

```
. reldist graph, olab(3000(1000)20000, format(%7.0g) grid) ///
```

```
>          yolab(3000(1000)20000, format(%7.0g) grid angle(0)) ///
```

```
>          ciopts(fc(%50) lc(%0))
```

# Relative density





# Relative density



```
. reldist pdf earnings [pw=wt], by(female) histogram notable
```

Relative density	Number of obs	=	100,000
F1: female = 1	Comparison obs	=	44,628
F0: female = 0	Reference obs	=	55,372
	Bandwidth	=	.02515569

```
. reldist graph, olab(3000(1000)20000, format(%7.0g) grid) ///  
>      ciopts(fc(%50) lc(%0))
```

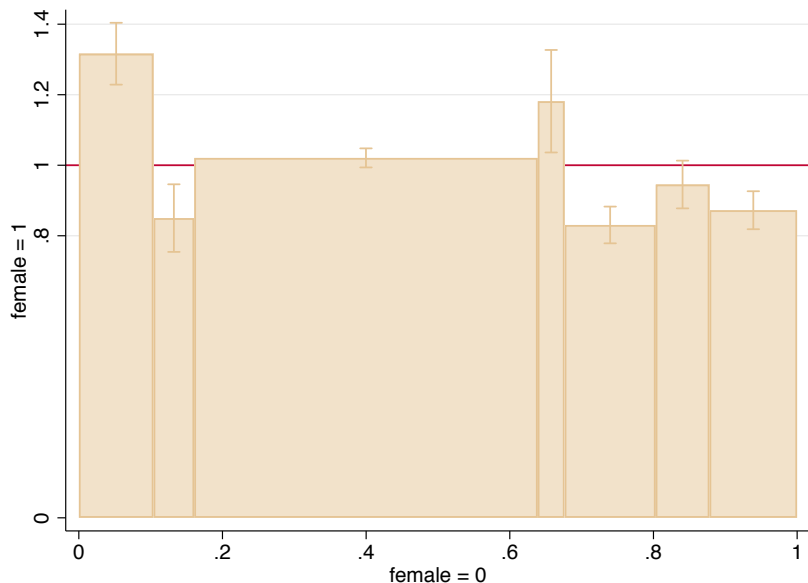
# Relative polarization

```
. reldist mrp earnings [pw=wgt], by(female) multiplicative
```

Median relative polarization	Number of obs	=	100,000
F1: female = 1	Comparison obs	=	44,628
F0: female = 0	Reference obs	=	55,372
Adjustment: location (mult)			

earnings	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
MRP	-.0465722	.0079613	-5.85	0.000	-.0621763	-.0309682
LRP	-.0033018	.0148662	-0.22	0.824	-.0324393	.0258358
URP	-.0898427	.0110417	-8.14	0.000	-.1114843	-.0682012

## Difference in covariates: education



## Difference in covariates: education



```
. reldist histogram educyrs [pw=wgt], by(female) categorical
```

Relative histogram

Number of obs

= 100,000

F1: female = 1

Comparison obs

= 44,628

F0: female = 0

Reference obs

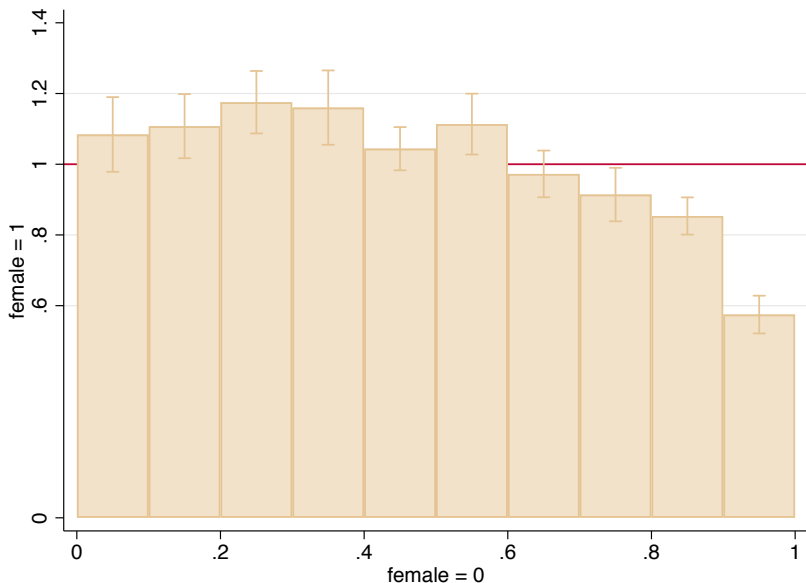
= 55,372

educyrs	Coef.	Std. Err.	[95% Conf. Interval]	
educyrs				
7	1.316267	.0447324	1.228592	1.403942
11	.8500557	.0489017	.754209	.9459024
12	1.020779	.0137853	.9937596	1.047798
13	1.181543	.0741483	1.036213	1.326873
14	.8305811	.0265873	.7784703	.8826918
15	.9453244	.0345518	.8776033	1.013045
17	.8723796	.0274635	.8185515	.9262076

(evaluation grid stored in e(at))

```
. reldist graph
```

## Difference in covariates: tenure



## Difference in covariates: tenure



```
. reldist histogram tenure [pw=wt], by(female)
```

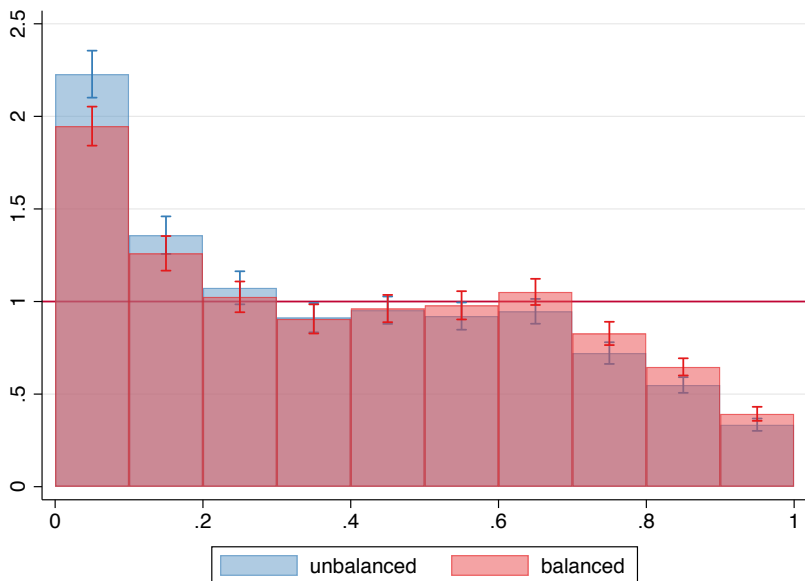
Relative histogram	Number of obs	=	100,000
F1: female = 1	Comparison obs	=	44,628
F0: female = 0	Reference obs	=	55,372

tenure	Coef.	Std. Err.	[95% Conf. Interval]	
h1	1.084155	.053922	.9784687	1.189842
h2	1.107638	.0462447	1.016999	1.198277
h3	1.175377	.0450791	1.087022	1.263731
h4	1.160171	.053622	1.055073	1.26527
h5	1.04392	.0311894	.9827894	1.105051
h6	1.113525	.043905	1.027472	1.199578
h7	.9726401	.0337204	.9065484	1.038732
h8	.9141628	.0385788	.8385488	.9897768
h9	.8535668	.0268357	.8009691	.9061645
h10	.5748437	.0272384	.5214568	.6282306

(evaluation grid stored in e(at))

```
. reldist graph
```

# Covariate balancing



# Covariate balancing



```
. reldist histogram earnings [pw=wgt], by(female)
(output omitted)
. estimates store unbalanced
. reldist histogram earnings [pw=wgt], by(female) ///
> balance(eb:i.educyrs c.tenure##c.tenure)
```

```
Relative histogram      Number of obs      =      100,000
F1: female = 1          Comparison obs      =      44,628
F0: female = 0          Reference obs       =      55,372
Balancing of F1
method = eb
i.educyrs tenure c.tenure#c.tenure
```

earnings	Coef.	Std. Err.	[95% Conf. Interval]	
h1	1.947315	.0537421	1.841982	2.052649
h2	1.26018	.047665	1.166757	1.353602
h3	1.025128	.0424014	.942022	1.108235
h4	.9059489	.0401832	.8271904	.9847074
h5	.9619829	.0375585	.8883687	1.035597
h6	.9794557	.0389595	.9030956	1.055816
h7	1.051987	.0360187	.9813911	1.122584
h8	.8276325	.0320939	.7647288	.8905361
h9	.6469504	.0236423	.6006119	.693289
h10	.3934189	.019337	.3555186	.4313191

```
(evaluation grid stored in e(at))
. estimates store balanced
. coefplot unbalanced balanced, at nooffset citop cirecast(rcap) ///
> recast(bar) barwidth(0.1) color(%50) ylabel(0(.5)2.5) yline(1)
```



# Covariate balancing

```
. reldist summarize earnings [pw=wgt], by(female) stat(mean med)
```

Relative ranks	Number of obs	=	100,000
F1: female = 1	Comparison obs	=	44,628
F0: female = 0	Reference obs	=	55,372

earnings	Coef.	Std. Err.	[95% Conf. Interval]	
mean	.3756438	.0034384	.3689046	.382383
median	.3348484	.0066729	.3217696	.3479273

```
. reldist summarize earnings [pw=wgt], by(female) stat(mean med) ///  
> balance(eb:i.educyrs c.tenure##c.tenure)
```

Relative ranks	Number of obs	=	100,000
F1: female = 1	Comparison obs	=	44,628
F0: female = 0	Reference obs	=	55,372

```
Balancing of F1  
method = eb  
i.educyrs tenure c.tenure#c.tenure
```

earnings	Coef.	Std. Err.	[95% Conf. Interval]	
mean	.4040611	.0027288	.3987127	.4094096
median	.3854737	.0057611	.3741821	.3967654

- 1 Introduction
- 2 Theory and estimation
- 3 The `reldist` command
- 4 Spinoff: a general command for the analysis of distributions

# Analysis of univariate distributions

- After deriving the equations and implementing `reldist`, I realized that I had all the building blocks in front of me for putting together a general command for the analysis of (univariate) distributions (summary statistics, density, quantile function, inequality measures, etc.).
- This may not seem very exciting.
- After all, many official (`mean`, `proportion`, `ci`, `summarize`, `tabstat`, `pctile`, `cumul`, `kdensity`, `histogram`, etc.) and user-written commands (`catplot`, `cdfplot`, `distplot`, `fre`, `kdens`, `lorenz`, `pshare`, `glcurve`, `svylorenz`, `roblast`, etc.) are available.

# Analysis of univariate distributions

- But it is!
- All these statistics can be combined in a general framework based on influence functions. This means that you get `svy`-compatible standard errors for everything (as well as covariances between any kind of statistic).
- Covariate balancing/standardization can easily be integrated in a general way.
- RIFs (recentered influence functions) are available for everything and can be used in further analysis, e.g. in RIF regressions or RIF decompositions.

## Estimation

### Scalar summary statistics

```
dstat [summarize] [(stats)] varlist [ (stats) varlist ... ] [if] [in] [weight] [, options ]
```

### Distribution functions

```
dstat subcmd varlist [if] [in] [weight] [, options ]
```

where *subcmd* is

<u>d</u> ensity	density function
<u>h</u> istogram	histogram
<u>p</u> roportion	proportions or totals
<u>c</u> df	cumulative distribution
<u>q</u> uantile	quantile function
<u>l</u> orenz	lorenz curve
<u>s</u> hare	percentile shares

*varlist* may contain factor variables; see `fvvarlist`.  
`fweights`, `pweights`, and `iweights` are allowed; see `weight`.

## Postestimation

### Replay results

```
dstat [, reporting_options ]
```

### Draw graph

```
dstat graph [, graph_options ]
```

### Obtain (recentered) influence functions

```
predict {stub* | newvar1 newvar2 ...} [if] [in] [, predict_options ]
```

stats	Description
Points in the distribution	
<code>quantile(p)</code>	$p/100$ -quantile; $p$ in $[0,100]$
<code>p(p)</code>	alias for <code>quantile()</code>
<code>density(x)</code>	kernel density at value $x$
<code>hist(x1,x2)</code>	histogram density of data within $[x1,x2]$
<code>cdf(x)</code>	cumulative distribution (CDF) at value $x$
<code>cdfm(x)</code>	mid-adjusted CDF at value $x$
<code>prop(x)</code>	proportion of data equal to value $x$
<code>prop(x1,x2)</code>	proportion of data within $[x1,x2]$
<code>pct(x)</code>	percent of data equal to value $x$
<code>pct(x1,x2)</code>	percent of data within $[x1,x2]$
<code>freq(x)</code>	frequency of data equal to value $x$
<code>freq(x1,x2)</code>	frequency of data within $[x1,x2]$
Location measures	
<code>mean</code>	arithmetic mean
<code>gmean</code>	geometric mean (data must be positive)
<code>hmean</code>	harmonic mean (data must be positive)
<code>trim([alpha])</code>	$\alpha$ trimmed mean; $\alpha$ in $[0,50]$ ; default is $\alpha=25$
<code>winsor([alpha])</code>	$\alpha$ winsorized mean; $\alpha$ in $[0,50]$ ; default is $\alpha=25$
<code>median</code>	median; equal to <code>q50</code>
<code>huber([p])</code>	Huber M estimate with gaussian efficiency $p$ in $[63.7,99.9]$ ; default is $p=95$
<code>biweight([p])</code>	biweight M estimate with gaussian efficiency $p$ in $[.01,99.9]$ ; default is $p=95$
<code>hl</code>	Hodges-Lehmann location measure (Hodges and Lehmann 1963)
Scale measures	
<code>sd([df])</code>	standard deviation; default is $df=1$
<code>variance([df])</code>	variance; default is $df=1$
<code>mse([t],[df])</code>	mean squared error from target $t$ ; default is $t=0$ and $df=0$
<code>smse([t],[df])</code>	square-root of mean squared error; default is $t=0$ and $df=0$
<code>iqr([p1,p2])</code>	interquantile range; default is <code>iqr(25,75)</code> (interquartile range)
<code>iqrn</code>	rescaled interquartile range
<code>mad([l],[t])</code>	median (or mean if $l!=0$ ) absolute deviation from the median (or mean if $t!=0$ )
<code>madn([l],[t])</code>	normalized MAD; equal to $1/\text{invnormal}(0.75) * \text{mad}$ or $\sqrt{\pi/2} * \text{mad}$
<code>mae([l],[t])</code>	median (or mean if $l!=0$ ) absolute error from target $t$ ; default is $t=0$
<code>maen([l],[t])</code>	normalized MAE; equal to $1/\text{invnormal}(0.75) * \text{mae}$ or $\sqrt{\pi/2} * \text{mae}$
<code>md</code>	mean absolute pairwise difference; equal to $2 * \text{mean} * \text{gini}$
<code>mdn</code>	normalized mean absolute pairwise difference; equal to $\sqrt{\pi/2} * \text{md}$
<code>mscale([bp])</code>	M estimate of scale with breakdown point $bp$ in $[1,50]$ ; default is $bp=50$
<code>qn</code>	Qn scale coefficient (Rousseeuw and Croux 1993)
Skewness measures	
<code>skewness</code>	skewness
<code>qskew([alpha])</code>	quantile skewness measure (Hinkley 1975); $\alpha$ in $[0,50]$ ; default is $\alpha=25$
<code>mc</code>	medcouple (Brys et al. 2004)

Kurtosis measures	
<b>kurtosis</b>	kurtosis
<b>qw[<i>(alpha)</i>]</b>	quantile tail weight measure; <i>alpha</i> in [0,50]; default is <i>alpha</i> =25
<b>lw[<i>(alpha)</i>]</b>	left quantile tail weight measure; <i>alpha</i> in [0,50]; default is <i>alpha</i> =25
<b>rw[<i>(alpha)</i>]</b>	right quantile tail weight measure; <i>alpha</i> in [0,50]; default is <i>alpha</i> =25
<b>lmc</b>	left medcouple tail weight measure (Brys et al. 2006)
<b>rmc</b>	right medcouple tail weight measure (Brys et al. 2006)
Inequality measures	
<b>gini[<i>(df)</i>]</b>	Gini coefficient; <i>df</i> applies small-sample adjustment; default is <i>df</i> =0
<b>agini[<i>(df)</i>]</b>	absolute Gini coefficient
<b>mlc</b>	mean log deviation; equal to <b>ge(0)</b>
<b>thell</b>	Theil index; equal to <b>ge(1)</b>
<b>cv[<i>(df)</i>]</b>	coefficient of variation; default is <i>df</i> =1; <b>cv(0)</b> = <b>sqrt(2*ge(1))</b>
<b>ge[<i>(alpha)</i>]</b>	generalized entropy (Shorrocks 1980) with parameter <i>alpha</i>
<b>atkinson[<i>(epsilon)</i>]</b>	Atkinson index with parameter <i>epsilon</i> =0; default is <i>epsilon</i> =1
<b>lvar[<i>(df)</i>]</b>	logarithmic variance; <i>df</i> applies small-sample adjustment; default is <i>df</i> =1
<b>vlog[<i>(df)</i>]</b>	variance of logarithm; <i>df</i> applies small-sample adjustment; default is <i>df</i> =1
<b>top[<i>(p)</i>]</b>	outcome share of top <i>p</i> percent; default is <i>p</i> =10
<b>bottom[<i>(p)</i>]</b>	outcome share of bottom <i>p</i> percent; default is <i>p</i> =40
<b>mid[<i>(p1,p2)</i>]</b>	outcome share of mid <i>p1</i> to <i>p2</i> percent; default is <i>p1</i> =40 and <i>p2</i> =90
<b>palma</b>	palma ratio; equal to <b>top/bottom</b> or <b>sratio(40,90)</b>
<b>qratio[<i>(p1,p2)</i>]</b>	quantile ratio $q(p2)/q(p1)$ ; default is <i>p1</i> =10 and <i>p2</i> =90
<b>sratio[<i>(u1,l2)</i>]</b>	percentile share ratio; default is <i>u1</i> =10 and <i>l2</i> =90
<b>sratio[<i>(l1,u1,l2,u2)</i>]</b>	percentile share ratio; default is <i>l1</i> =0, <i>u1</i> =10, <i>l2</i> =90, <i>u2</i> =100
<b>*lorenz(<i>p</i>)</b>	Lorenz ordinate, <i>p</i> in [0,100]; prefix * is empty for default, <b>g</b> for generalized, <b>t</b> for total, <b>a</b> for absolute, <b>e</b> for equality gap
<b>*share(<i>p1,p2</i>)</b>	percentile share, <i>p1</i> and <i>p2</i> in [0,100]; prefix * is empty for default, <b>d</b> for density, <b>g</b> for generalized, <b>t</b> for total, <b>a</b> for average
Concentration measures	
<b>gci[<i>(zvar,df)</i>]</b>	Gini concentration index; <i>zvar</i> specifies the sort variable; <i>df</i> applies small-sample adjustment; default is <i>df</i> =0
<b>gci[<i>(df)</i>]</b>	<b>gci</b> using sort variable from option <i>zvar</i> ()
<b>aci[<i>(zvar,df)</i>]</b>	absolute Gini concentration index; <i>zvar</i> and <i>df</i> are as for <b>gci</b>
<b>aci[<i>(df)</i>]</b>	<b>aci</b> using sort variable from option <i>zvar</i> ()
<b>*ccurve(<i>p,zvar</i>)</b>	concentration curve ordinate, <i>p</i> in [0,100]; prefix * is empty for default, <b>g</b> for generalized, <b>t</b> for total, <b>a</b> for absolute, <b>e</b> for equality gap
<b>*cshare(<i>p1,p2,zvar</i>)</b>	concentration share, <i>p1</i> and <i>p2</i> in [0,100]; prefix * is empty for default, <b>d</b> for density, <b>g</b> for generalized, <b>t</b> for total, <b>a</b> for average
Poverty measures	
<b>watts[<i>(pline)</i>]</b>	Watts index (see, e.g., Saisana 2014); <i>pline</i> specifies the poverty line(s) > 0; <i>pline</i> can be <i>varname</i> or #; the default is as set by option <b>pline()</b>
<b>fgt[<i>(a,pline)</i>]</b>	Foster-Greer-Thorbecke index with <i>a</i> >0 (Foster et al. 1984, 2010); default is <i>a</i> =0 (headcount ratio); <i>pline</i> specifies the poverty line(s) > 0; <i>pline</i> can be <i>varname</i> or #; the default is as set by option <b>pline()</b>

# Example

```
. dstat (mean gmean med sd Gini MLD Theil Palma) earnings [pw=wgt], over(female)
```

```
Summary statistics          Number of obs   =   100,000
```

```
0: female = 0
```

```
1: female = 1
```

earnings	Coef.	Std. Err.	[95% Conf. Interval]	
0				
mean	7964.767	32.99754	7900.093	8029.442
gmean	7231.028	23.98644	7184.015	7278.041
med	6803	27.13438	6749.817	6856.183
sd	4539.07	102.4153	4338.337	4739.803
Gini	.2433624	.0019915	.239459	.2472657
MLD	.0966465	.001718	.0932792	.1000138
Theil	.1137077	.0027248	.1083671	.1190484
Palma	.8660138	.00943	.8475311	.8844965
1				
mean	6515.329	24.85582	6466.611	6564.046
gmean	6082.104	18.92963	6045.003	6119.206
med	5893	26.14387	5841.758	5944.242
sd	2897.98	78.86047	2743.415	3052.546
Gini	.2061163	.001989	.2022179	.2100147
MLD	.0688069	.0015461	.0657765	.0718373
Theil	.076873	.0023677	.0722324	.0815136
Palma	.7110416	.0084675	.6944454	.7276379



# Installation

- `reldist` requires the latest version of `moremata`. To install both packages, type

```
. ssc install reldist, replace
```

```
. ssc install moremata, replace
```

Or install from GitHub: <http://github.com/benjann/reldist>

- `dstat` should become available on GitHub and SSC soon; check <http://github.com/benjann/dstat> in some weeks.

# References

- Handcock, M.S., M. Morris (1998). Relative Distribution Methods. *Sociological Methodology* 28: 53-97.
- Handcock, M.S., M. Morris (1999). *Relative Distribution Methods in the Social Sciences*. New York: Springer.
- Jann, B. (2020a). Influence functions continued. A framework for estimating standard errors in reweighting, matching, and regression adjustment. University of Bern Social Sciences Working Papers 35. Available from <https://ideas.repec.org/p/bss/wpaper/35.html>.
- Jann, B. (2020b). Relative distribution analysis in Stata. University of Bern Social Sciences Working Papers 37. Available from <http://ideas.repec.org/p/bss/wpaper/37.html>.
- Morris, M., A.D. Bernhardt, M.S. Handcock (1994). Economic Inequality: New Methods for New Trends. *American Sociological Review* 59: 205-219.